**CMSC 35900 (Spring 2008) Learning Theory** **Lecture: 13a**

# (Exponentiated) Stochastic Gradient Descent for L1 Constrained Problems

*Instructors: Sham Kakade and Ambuj Tewari*

**Abstract**

This note is by Sham Kakade, Dean Foster, and Eyal Even-Dar. It is intended as an introductory piece on solving $L_1$ constrained problems with online methods.

Convex optimization problems with $L_1$ constraints frequently underly solving such tasks as feature selection problems and obtaining sparse representations. This note shows that the exponentiated gradient algorithm (of Kivinen and Warmuth (1997)) when used as a stochastic gradient descent algorithm is quite effective as an optimization tool under *general* convex loss functions — requiring a number of gradient steps that is logarithmic in the number of dimensions under mild assumptions. In particular, for supervised learning problems in which we desire to approximately minimize some general convex loss (including the square, logistic, hinge, or absolute loss) in the presence of many irrelevant features, this algorithm is efficient — with a *sample complexity* that is only logarithmic in the total number of features and a *computational complexity* that is only linear in the total number of features (ignoring log factors).

## 1 Introduction

In the past decade, there has been a flourishing interest in *sparse* methods — methods which favor and discover a small number of preferred elements relevant to some objective. Abstractly, the underlying optimization problem is to minimize some error subject to using only a small *number* of relevant elements — this latter constraint essentially uses the $L_0$ norm (the $L_0$ norm of a vector is the number of non-zero components). As this optimization problem is usually computationally intractable, a growing body of work has examined using the *convex* $L_1$ norm in lieu of the $L_0$ norm. The $L_1$ norm penalizes by the sum of the absolute values of the parameters and is the closest convex approximation to the $L_0$ norm (among the $L_p$ norms).

In the feature selection problem in machine learning and statistics, there is assumed to be a small set of features, among some larger set of irrelevant features, able to accurately approximate some target concept. Here, much is understood on the statistical issues related to finding relevant subsets of features (Foster and George (1994); Donoho and Johnstone (1994); Miller (2002)). As this subset selection problem is computationally intractable (Welch (1982)), a growing body of work has examined using the $L_1$ norm as a regularizer on the parameters, in order to approximately enforce sparsity (e.g. the LASSO algorithm of Tibshirani (1996)). A growing body of work shows that algorithms which use this penalization have feature selection properties (e.g. Ng (2004)).

In the signal processing community, one increasingly popular assertion is that the observed signal (e.g. image, speech signal, etc.) is composed of (a transformation of) a small number of relevant components plus noise. The goal is to recover the underlying signal. The matching pursuit algorithm of Mallat and Zhang (1993) attempts this by greedily choosing a small *number* of components which suffices to (approximately) reconstruct the signal — in essence, it is an $L_0$ style of optimization. Subsequently, Chen et al. (1999) introduced the basis pursuit algorithm, which approximately enforces sparsity in the signal reconstruction using an $L_1$ penalty. In this community, there is a growing literature showing how the $L_1$ solution can be a good approximation to the true underlying sparse solution (e.g. Donoho and Elad (2002); Candes and Tao (2006)).

**The Optimization Problem:** The core underlying optimization problem in these settings is:

$$\max_{w} \quad c(w) \tag{1}$$

$$\text{such that} \quad ||w||_1 \le F_1 \tag{2}$$

where $c(\cdot)$ is convex, $w \in \mathbb{R}^d$, $||w||_1 \equiv \sum_i |w_i|$ is the $L_1$ norm, and $F_1$ is a constant.

For example, in supervised learning using linear predictors, the loss function is often $\mathbb{E}[\ell(w \cdot x, y)]$. Here $x \in \mathbb{R}^d$ and the expectation is with respect to some underlying distribution. Often $\ell(z, y)$ is a convex loss in $z$, such as the square loss $(w \cdot x - y)^2$, the logistic loss (for binary or multi-class regression), the absolute loss $|w \cdot x - y|$, or the hinge loss (used in SVMs). These have all been considered with $L_1$ regularization. For the square loss case, the Least Angle Regression algorithm of Efron et al. (2002) is one of the fastest *exact* methods — having a runtime that is (theoretically) cubic in the total number of features.

For sufficiently large data sets, we desire fast *approximate* methods — conventional algorithms which loop over the dataset at least once before making *any* progress can be very inefficient. For this reason, among others, the stochastic gradient descent algorithm — which follows noisy gradient estimates (e.g. by taking the gradient at one training example) — has become a popular general purpose optimization algorithm.

## 2 Online Convex Programming and L1 vs. L2 Constraints

The *online convex programming problem* of Zinkevich (2003) consists of a convex feasible set $F \subset \mathbb{R}^d$ and a sequence $\{c^1, c^2, \ldots c^T\}$ where each $c^t : F \to \mathbb{R}$ is a convex function. At each timestep $t$, an online convex programming algorithm selects a vector $w^t \in F$. Subsequently, the algorithm incurs the cost $c^t(w^t)$. Importantly, no statistical assumptions on the sequence of convex functions are made — they should be thought of as an arbitrary sequence unknown apriori to the algorithm.

If algorithm $A$ uses the sequence of decisions $\{w^1, \ldots, w^T\}$ on the sequence $\{c^1, \ldots, c^T\}$, then $A$ has regret, at time $T$ in comparison to the best constant decision, defined as:

$$R_T(A) = \frac{1}{T}\left(\sum_{t=1}^{T} c^t(w^t) - \inf_{w \in F} \sum_{t=1}^{T} c^t(w)\right)$$

In this Section, we are interested in algorithms with little regret.

In Zinkevich (2003), the cost function $c^t$ is revealed to the algorithm only after the decision $w^t$ is chosen. Hence at time $t$, the algorithm has knowledge of the previous functions, $\{c^1(\cdot), \ldots, c^{t-1}(\cdot)\}$. In the bandit setting of Flaxman et al. (2005), only the costs incurred are revealed to the algorithm, so at time $t$, only $\{c^1(w^1), \ldots, c^{t-1}(w^{t-1})\}$ is known. In our setting, instead of the costs being revealed, noisy estimates of the gradient (or simply a sub-gradient[1]) are revealed. Specifically, denoting the gradient of $c^t$ at $w^t$ by $\nabla c^t(w^t)$ and the noisy version by $\widehat{\nabla} c^t(w^t)$, the algorithm only knows $\{\widehat{\nabla} c^1(w^1), \ldots, \widehat{\nabla} c^{t-1}(w^{t-1})\}$ at time $t$. This assumption is parsimonious with stochastic gradient optimization procedures, which are considered in the next Section.

We make the following assumptions, which depend on whether we are using $L_1$ or $L_2$ norms.

1. The constraints on the shape of the feasible set $F \subset \mathbb{R}^d$ are as follows:

   **L2:** Assume that $F$ is convex, closed, non-empty, and bounded. In particular, there exists a constant $F_2$ such for all $w \in F$,
   $$||w||_2 \leq F_2$$
   where $||w||_2 \equiv \sqrt{\sum_i w_i^2}$.

   **L1:** Assume that $F$ is a (scaled) simplex. The scale of the simplex, called $F_1$, describes its size, i.e. for all $w \in F$ we have that $w_i \geq 0$ and
   $$||w||_1 = F_1$$
   .

2. The estimated gradient is unbiased condition on the history $\mathcal{H}^{t-1}$, i.e.
   $$\mathbb{E}[\widehat{\nabla} c^t(w^t) | \mathcal{H}^{t-1}] = \nabla c^t(w^t)$$
   where $\mathcal{H}^{t-1}$ denotes the outcomes up to and including time $t-1$.

[1] If the gradient of any $c_t$ does not exist, a sub-gradient suffices with no change in the results.

3. Assume the estimated and exact gradients are bounded. In particular,

   **L2:** With probability one, for all $w \in F$ and times $t$, that both

   $$||\nabla c^t(w)||_2 \leq G_2, \quad \text{and} \quad ||\widehat{\nabla} c^t(w)||_2 \leq G_2$$

   **L1:** With probability one, for all $w \in F$ and times $t$, that both

   $$||\nabla c^t(w)||_\infty \leq G_\infty, \quad \text{and} \quad ||\widehat{\nabla} c^t(w)||_\infty \leq G_\infty$$

   where $||w||_\infty = \max_i w_i$ is the $L_\infty$ norm.

Assumptions 2 and 3 are common to most stochastic approximation settings. In particular, it is important that Assumption 2 be *conditionally* unbiased, as noise typically depends on the current parameter $w^t$ (which in turn depends on the history).

The crucial distinction between using $G_2$ and $G_\infty$ is that often $G_2$ often has an implicit dependence on the number of dimensions while $G_\infty$ often has no dependence on the number of dimensions (as it is a bound on *maximum*, rather than a sum, over all the dimensions). Furthermore, $F_1$ should be thought of as a parameter that is set to be large enough to include the weights of all the relevant parameters, which could be far less than the total number of dimensions. For example, if we believe there are $p$ relevant dimensions (out of the $d$) with weight less than 1, then an appropriate setting is $F_1 = p$ (Candes and Tao (2006) makes this precise). Hence, in this setting, the product $F_1 G_\infty$ should be viewed as $O(p)$ while $F_2 G_2$ has a strong dimensionality dependence (more like $O(d)$).

All proofs in this Section are provided in the Appendix.

## 2.1 Gradient Descent with L2 Constraints

Here, we reconsider the case in Zinkevich (2003), extending the results for when only noisy gradients are provided. We work with the $L_2$ assumptions.

Define the Stochastic Gradient descent algorithm (SG) with fixed learning rate $\eta$ is as follows: at $t = 1$, select any $w^1 \in F$, and update the decision as follows

$$w^{t+1} = \text{Proj}_F[w^t - \eta \widehat{\nabla} c^t(w^t)]$$

where $\text{Proj}_F[w]$ is the projection of $w$ back into $F$, i.e. it is the closest point (under the $L_2$ norm) in $F$ to $w$. Hence, $w^{t+1} \in F$.

The following bound was provided in Zinkevich (2003) for the noiseless case.[2] We also provide a result for the noisy case.

**Theorem 2.1.** *Set* $\eta = \frac{2F_2}{G_2}\sqrt{\frac{1}{T}}$.

- *(Noiseless Case) (from Zinkevich (2003)) Assuming the gradient estimates are noise free, the regret of SG at time $T$ is bounded as follows:*

  $$R_T(SG) \leq 2F_2 \, G_2 \sqrt{\frac{1}{T}}$$

- *(Noisy Case) If the estimates are noisy, then with probability greater than $1 - \delta$, the regret of SG at time $T$ is bounded as follows:*

  $$R_T(SG) \leq 2F_2 \, G_2 \left( \sqrt{\frac{1}{T}} + \sqrt{\frac{8}{T} \log \frac{1}{\delta}} \right)$$

As discussed earlier, in many cases $F_2 G_2$ implicitly depends on the number of dimensions.

[2] We state this Theorem rather differently than in Zinkevich (2003), as we use a learning rate that is appropriately scaled based on $F_2$ and $G_2$.

## 2.2 Stochastic Multiplicative Gradient Descent with L1 Constraints

Assume the decision space $F$ is a (scaled) $d$-dimensional simplex. We define the Stochastic Multiplicative Gradient descent algorithm (SMG) as follows: at time $t = 1$, choose $w^1$ as the center point of the simplex, namely $w_i^1 = \frac{F_1}{d}$, and update the coordinates of the parameters according to:

$$(\forall i) \quad w_i^{t+1} = w_i^t \left(1 - \eta[\widehat{\nabla} c^t(w^t)]_i + \eta Z\right) \quad \text{where} \quad Z = \frac{w^t \cdot \widehat{\nabla} c^t(w^t)}{F_1} \tag{3}$$

where $[\widehat{\nabla} c^t(w^t)]_i$ is the $i$-th component of the noisy gradient, and $w \cdot w'$ denotes the inner product $\sum_{i=1}^d w_i w_i'$. Here, the subtraction by $Z$ serves as a form of projection, so that $w^{t+1} \in F$ — it is straight forward to verify that $w^{t+1} \in F$ if $w^t \in F$ and if $\eta$ is sufficiently small (such that the update prevents any coordinate from becoming negative).

This SMG update rule is essentially a first order (in $\eta$) approximation of the exponentiated gradient update rule of Kivinen and Warmuth (1997). This form is chosen as it makes the connection to gradient descent more explicit. In particular, the following Proposition shows that the multiplicative gradient descent algorithm has a positive inner product with the gradient descent direction (so it has a negative dot product with the gradient). The importance of this is that it shows that the SMG really points in the direction of the gradient — suggesting it is reasonable to consider in a stochastic gradient optimization method, which is done in the next Section.

**Proposition 2.2.** *If the gradient estimates are noiseless, the multiplicative gradient descent algorithm has a positive inner with the gradient descent direction, i.e.*

$$(w^{t+1} - w^t) \cdot \nabla c^t(w^t) \leq 0$$

*Proof.* For notational convenience, let $\nabla^t = \nabla c^t(w^t)$, and $\nabla_i^t = [\widehat{\nabla} c^t(w^t)]_i$. Furthermore, without loss of generality, consider the case where $F_1 = 1$, so $w$ can be viewed as probability distributions (if $F_1 \neq 1$ then the parameters can be rescaled). We have:

$$(w^{t+1} - w^t) \cdot \nabla^t = \eta \sum_i \left(-w_i^t \nabla_i^t + (w^t \cdot \nabla^t) w_i^t\right) \nabla_i^t = -\eta \left(\sum_i w_i^t (\nabla_i^t)^2 - (\sum_i w_i^t \nabla_i^t)^2\right)$$

The proof is completed by noting that the last term takes the form of a (negative) variance. □

We now state one of our main results. Note that it is uses the dual norms $L_1/L_\infty$ rather than $L_2/L_2$.

**Theorem 2.3.** *Assume that $T \geq 16 \log d$. Set the learning rate as $\eta = \frac{1}{2F_1 G_\infty}\sqrt{\frac{\log d}{T}}$. We have:*

- *(Noiseless Case) Assuming the gradient estimates are noise free, the $SMG$ has regret at time $T$ bounded as:*

$$R_T(SMG) \leq 4F_1 G_\infty \sqrt{\frac{\log d}{T}}$$

- *(Noisy Case) If the estimates are noisy, then with probability greater than $1 - \delta$, the $SMG$ has regret at time $T$ bounded as:*

$$R_T(SMG) \leq 4F_1 G_\infty \left(\sqrt{\frac{\log d}{T}} + \sqrt{\frac{2}{T} \log \frac{1}{\delta}}\right)$$

Hence, when $F_1 G_\infty$ is $O(p)$ (where $p$ is the number of relevant dimensions), this bound is only logarithmic in the total number of dimensions. This point is returned to in Section 4, where we consider the supervised learning setting.

# 3 Convex Optimization with L1 Constraints

Returning to the optimization problem discussed in the Introduction (see Equation 1), consider:

$$\begin{aligned} \max_{w} \quad & c(w) \\ \text{such that} \quad & ||w||_1 = F_1,\ w \geq 0 \end{aligned}$$

where $c(w)$ is convex. Although this optimization problem is slightly different from Equation 1, under simple transformations they are essentially equivalent.[3]

For this optimization problem, assume noisy estimates, $\widehat{\nabla} c$, of the true gradient $\nabla c$ can be obtained. We desire an algorithm that has low computational cost and only requires a few number of gradient estimates. The reason for desiring a small number of gradient estimates is that often times a gradient estimate corresponds to some costly Monte Carlo sample, such as the in the supervised learning setting in following Section, where a gradient sample corresponds to a obtaining a training example.

The SMG algorithm for this case is simply to perform the update at time $t$ using the gradient estimate obtained at time $t$. Note that Proposition 2.2 implies that the inner product between the SMG and the gradient descent direction is non-negative. However, this does not necessarily imply that this inner product will be large,[4] so standard stochastic approximation techniques do not imply that SMG leads to an effective solution. However, the following Corollary, shows that averaging parameters, does converge quickly to the optimal loss. A similar point was noted in Cesa-Bianchi et al. (2004).

**Corollary 3.1.** *Define the average decision to be $\overline{w}^T = \frac{1}{T}\sum_{t=1}^{T} w^t$. With the learning rate and conditions specified in Theorem 2.3, then with probability greater than $1 - \delta$, the following performance bound holds after $T$ updates:*

$$c(\overline{w}^T) \leq c(w^*) + 4F_1 G_\infty \left( \sqrt{\frac{\log d}{T}} + \sqrt{\frac{2}{T} \log \frac{1}{\delta}} \right)$$

*where $w^*$ is a solution to the optimization problem. Furthermore, if we continue running the algorithm, with the same constant setting of the learning rate, then almost surely,*

$$\limsup_{\tau \to \infty} c(\overline{w}^\tau) \leq c(w^*) + 4F_1 G_\infty \sqrt{\frac{\log d}{T}}$$

*Proof.* The first part of the proof follows by noting that convexity implies that $c(\overline{w}^T) \leq \frac{1}{T}\sum_{t=1}^{T} c(w^t)$ and the latter quantity is bounded in Theorem 2.3. The asymptotic statement follows using a martingale convergence theorem in Lemma 5.1. □

Note that this bound only weakly depends on the total number of dimensions $d$, when $F_1 G_\infty$ is roughly the size of the number of relevant dimensions. Hence, if an $\epsilon$ (additive) approximation is desired, a number of gradient updates that is only logarithmic in $d$ and linear in $\frac{1}{\epsilon^2}$ suffices. Furthermore, if we assume each gradient update requires $O(d)$ computations (as $d$ parameters are altered with a single update), then the total computational time required is only linear in $d$ (up to log factors).

# 4 Learning, Generalization, and Feature Selection

Consider the standard supervised learning framework where we observe a training set $\{(x^1, y^1), \ldots, (x^T, y^T)\}$, where each $(x^t, y^t)$ is sampled independently from a fixed and unknown underlying distribution $\mathcal{X} \times \mathcal{Y}$. Using this training set, we are interested in obtaining a hypothesis $h$, a mapping from $\mathcal{X}$ to some decision space. Measure the quality of our hypothesis on an example $(x, y)$ with the loss function $\ell(h(x), y)$. We are interested in obtaining a classifier

[3] To allow negative weights introduce two positive variables $w_+$ and $w_-$, with $w = w_+^t - w_-^t$. If the decision space needs to include points interior points ($||w||_1 < F_1$), include a dummy dimension, where the cost has no dependence on this dummy variable $w_{d+1}$.

[4] When $w$ is close to the corners of the simplex this inner product could be arbitrarily small. See the proof of Proposition 2.2 and note when the quantity that is the variance could be small.

which has low expected risk, where risk(h) $= \mathbb{E}\ell(h(x), y)$, where the expectation is with respect to the underlying distribution.

We make the following additional assumptions:

1. The input space is $\mathcal{X} \subset \mathbb{R}^d$. For simplicity, assume $||x||_\infty \leq 1$ for all $x \in \mathcal{X}$.
2. The loss function depends only on $w \cdot x$, so the loss suffered on example $(x, y)$ is $\ell(w \cdot x, y)$.
3. The loss function $\ell(z, y)$ is a convex function of $z$, for all $y \in \mathcal{Y}$.
4. Assume $\left|\frac{\partial \ell(z,y)}{\partial z}\right| \leq G_\infty$ for all $y \in \mathcal{Y}$ and $z \in \{w \cdot x : w \in F, x \in \mathcal{X}\}$

Many widely used learning algorithms fall under these assumptions — including least squares regression, (binary and multi-class) logistic regression, and algorithms which use the hinge or absolute loss. Basis pursuit de-noising (Chen et al. (1999)) can also be viewed in this setting. Assumption 1 corresponds to a setting where $x$ is, say, a feature vector and each feature is binary or bounded in $[-1, 1]$. Furthermore, the constant $G_\infty$ often has no dimensionality dependence — for the hinge, logistic, and absolute loss, $G_\infty = 1$, and for the square loss $G_\infty$ is a constant (depending on the maximal prediction error).

We consider the performance of an algorithm which makes one pass over the training set.

**Corollary 4.1.** *Let $w^*$ be the minimizer of $risk(w)$ over $w \in F$. Assume SMG uses $\nabla \ell(w^t \cdot x^t, y^t)$ as the gradient estimate, with the learning rate specified in Theorem 2.3. Again let $\overline{w}^T = \frac{1}{T}\sum_{t=1}^{T} w^t$. With probability greater than $1 - \delta$, after $T$ updates, the risk is bounded as follows:*

$$\textit{risk}\left(\overline{w}^T\right) \leq \textit{risk}(w^*) + 4F_1 G_\infty \left(\sqrt{\frac{\log d}{T}} + \sqrt{\frac{2}{T}\log\frac{1}{\delta}}\right)$$

This bound has the same feature selection properties as discussed in Kivinen and Warmuth (1997); Ng (2004), yet it applies to general (and unbounded) convex loss functions. Again, if we view $F_1 G_\infty$ as being order the number of relevant features, then the number of samples required for a good generalization error scales only logarithmic with the total number of features.

# References

Candes, E. J. and Tao, T. (2006). The dantzig selector: statistical estimation when $p$ is much larger than $n$. *To appear, Annals of Statistics*.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

Donoho, D. and Elad, M. (2002). Optimally sparse representation in general (non-orthogonal) dictionaries via l1 minimization. *Proc. Nat. Aca. Sci.*, 100.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2002). Least angle regression. Technical report, Stanford University.

Flaxman, A., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: gradient descent without a gradient. *SODA*.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, 22:1947–1975.

Kivinen, J. and Warmuth, M. (1997). Exponentiated gradient versus gradient descent for linear predictors. *Journal of Information and Computation*, 132.

Mallat, S. and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415.

Miller, A. J. (2002). *Subset Selection in Regression (Second Edition)*. Chapman& Hall, London.

Ng, A. Y. (2004). Feature selection, l 1 vs. l 2 regularization, and rotational invariance. *ICML.*

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58.

Welch, W. J. (1982). Algorithmic complexity: Three np-hard problems in computational statistics. *J. Statist. Comput. Simul.*

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. *ICML.*

# 5 Appendix

Throughout, let $\nabla^t = \nabla c^t(w^t)$ and $\widehat{\nabla}^t = \widehat{\nabla} c^t(w^t)$ for convenience. We start with a useful Lemma.

**Lemma 5.1.** *For algorithm $A$, assume that*

$$|(\nabla^t - \widehat{\nabla}^t) \cdot (w^t - w)| \leq B$$

*holds with probability* $1$. *Then with probability greater than* $1 - \delta$,

$$R_T(A) \leq \frac{1}{T}\sum_{t=1}^{T} \widehat{\nabla}^t \cdot (w^t - w^*) + B\sqrt{\frac{2}{T}\log\frac{1}{\delta}} \tag{4}$$

*Proof.* A fundamental property of convexity is that $c^t(w) \leq \nabla^t \cdot (w - w^t) + c^t(w^t)$ for all $w \in F$. Let $w^*$ be a minimizer of $\sum_t^T c^t(w)$ (which exists since $F$ is closed and convex). We have

$$R_T(A) = \frac{1}{T}\sum_{t=1}^{T}\left(c^t(w^t) - c^t(w^*)\right) \leq \frac{1}{T}\sum_{t=1}^{T} \nabla^t \cdot (w^t - w^*)$$

Note that by assumption, $\mathbb{E}[\widehat{\nabla}^t | \mathcal{H}^{t-1}] = \nabla^t$, which implies that $M_t = \sum_{\tau=1}^{t}(\widehat{\nabla}^\tau - \nabla^\tau) \cdot (w^\tau - w^*)$ is a martingale with respect to the filtration $\{\mathcal{H}^{t-1}\}$ and that $\mathbb{E}[M_1] = 0$. Furthermore, since $|M_t - M_{t-1}| = |(\widehat{\nabla}^t - \nabla^t) \cdot (w^t - w^*)|$ is bounded by $B$, the Azuma-Hoeffding bound implies

$$\frac{1}{T}\sum_{t=1}^{T}(\nabla^t - \widehat{\nabla}^t) \cdot (w^t - w^*) \leq B\sqrt{\frac{2}{T}\log\frac{1}{\delta}}$$

Combining this with the previous bound completes the proof. □

In the noiseless case, $B$ is 0. In the noisy $L_2$ case, $B \leq 4F_2G_2$, and in the noisy $L_1$ case, $B \leq 4F_1G_\infty$ — using the definitions of $F_2$, $G_2$, $F_1$ and $G_\infty$. We proceed to bound the first term in Equation 4 for both the SG and the MSG algorithms.

Below, we essentially provide the proof in Zinkevich (2003) for the SG algorithm. Theorem 2.1 then directly falls from the following Lemma and the previous Lemma — choosing the learning rate which minimizes the upper bound.

**Lemma 5.2.** *(SG) With probability greater than* $1 - \delta$, *the decisions of SG algorithm satisfy:*

$$\frac{1}{T}\sum_{t=1}^{T} \widehat{\nabla}^t \cdot (w^t - w^*) \leq \frac{2}{\eta T}F_2^2 + \frac{\eta}{2}G_2^2$$

*Proof.* Again let $w^*$ be the minimizer of the average cost. Define the distance $D(w, w')$ as the square Euclidean distance $||w - w'||_2^2$. A fundamental property of projections into convex bodies is that for an arbitrary $w' \in \mathbb{R}^d$, $D(\text{Proj}_F[w'], w) \leq D(w', w)$ for all $w \in F$. Hence, using this and the definition of $D$,

$$\begin{aligned} D(w^t, w^*) - D(w^{t+1}, w^*) &\geq D(w^t, w^*) - D(w^t - \eta\widehat{\nabla}^t, w^*) \\ &= ||w^t - w^*||_2^2 - ||w^t - \eta\widehat{\nabla}^t - w^*||_2^d \\ &= 2\eta\widehat{\nabla}^t \cdot (w^t - w^*) - \eta^2||\widehat{\nabla}^t||_2^2 \end{aligned}$$

and so

$$\widehat{\nabla}^t \cdot (w^t - w^*) \leq \frac{1}{2\eta}(D(w^t, w^*) - D(w^{t+1}, w^*)) + \frac{\eta}{2}G_2^2$$

using the definition of $G_2$. Summing over $t$,

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{\nabla}^t \cdot (w^t - w^*) \leq \frac{1}{2\eta T}(D(w^1, w^*) - D(w^{T+1}, w^*)) + \frac{\eta}{2}G_2^2 \leq \frac{2}{\eta T}F_2^2 + \frac{\eta}{2}G_2^2$$

where the last step uses the fact that $D(w^1, w^*) \leq 4F_2^2$ and $D(w^{T+1}, w^*) \geq 0$. □

To complete the proof for the SMG algorithm (Theorem 2.3), we use the following Lemma along with Lemma 5.1. We must choose the learning specified in the Theorem and verify that the technical condition below is satisfied.

**Lemma 5.3.** *(SMG) Let $w^*$ be the minimizer of $\sum_t^T c^t(w)$. If $\eta \leq \frac{1}{8G_\infty}$, then the SMG algorithm's decisions remain in the feasible set and satisfy*

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{\nabla}^t \cdot (w^t - w^*) \leq \frac{1}{T\eta}F_1 \log d + 4\eta F_1 G_\infty^2$$

*with probability greater than $1 - \delta$.*

*Proof.* For now, assume $F_1 = 1$ (we rescale $F$ later). Interpret $w \in F$ as a probability distribution. Let us examine how the KL-distance changes with respect to $w^*$.

$$KL(w^*||w^t) - KL(w^*||w^{t+1}) = \sum_i w_i^* \log \frac{w_i^{t+1}}{w_i^t} = \sum_i w_i^* \log(1 - \eta\nabla_i^t + \eta w^t \cdot \nabla^t)$$

We lower bound the previous equation using that the log functions satisfies the lower bound $\log(1 + x) \geq 1 + x - x^2$ for $x \geq -1/4$, which can be verified through a Taylor expansion.

With our assumption on $\eta$, it follows that $|-\eta\nabla_i^t + \eta w^t \cdot \nabla^t|$ is bounded by $2\eta||\nabla c||_\infty \geq 1/4$. Hence, this implies that $w^{t+1}$ does not have any negative components, so the decisions are feasible. Furthermore, since the log lower bound is valid, we can continue bounding:

$$\begin{aligned}
&\geq \sum_i w_i^*(-\eta\nabla_i^t + \eta w^t \cdot \nabla^t) - \sum_i w_i^*(-\eta\nabla_i^t + \eta w^t \cdot \nabla^t)^2 \\
&= \eta\sum_i w_i^*(w^t \cdot \nabla^t - \nabla_i^t) - \eta^2\sum_i w_i^*(\nabla_i^t - w^t \cdot \nabla^t)^2 \\
&\geq \eta\nabla^t \cdot (w^t - w^*) - 4\eta^2 G_\infty^2
\end{aligned}$$

where the last step uses $\sum_i w_i^*(w^t \cdot \nabla^t) = (w^t \cdot \nabla^t)\sum_i w_i^* = w^t \cdot \nabla^t$ and the definition of $G_\infty$.

Rearranging leads to:

$$\nabla^t \cdot (w^t - w^*) \leq \frac{1}{\eta}(KL(w^*||w^t) - KL(w^*||w^{t+1})) + 4\eta G_\infty^2$$

which implies

$$\frac{1}{T}\sum_{t=1}^{T}\nabla^t \cdot (w^t - w^*) \leq \frac{1}{\eta T}(KL(w^*||w^1) - KL(w^*||w^T)) + 4\eta G_\infty^2$$

The proof is completed by noting that $KL(w^*||w^T) \geq 0$ and $KL(w^*||w^1) \leq \log d$, since $w^1$ is uniform. Also, the feasible set must be rescaled to $F_1$ size. □