ℓ

− /

− −

http://jmlr.org/papers/v22/19-477.html

–

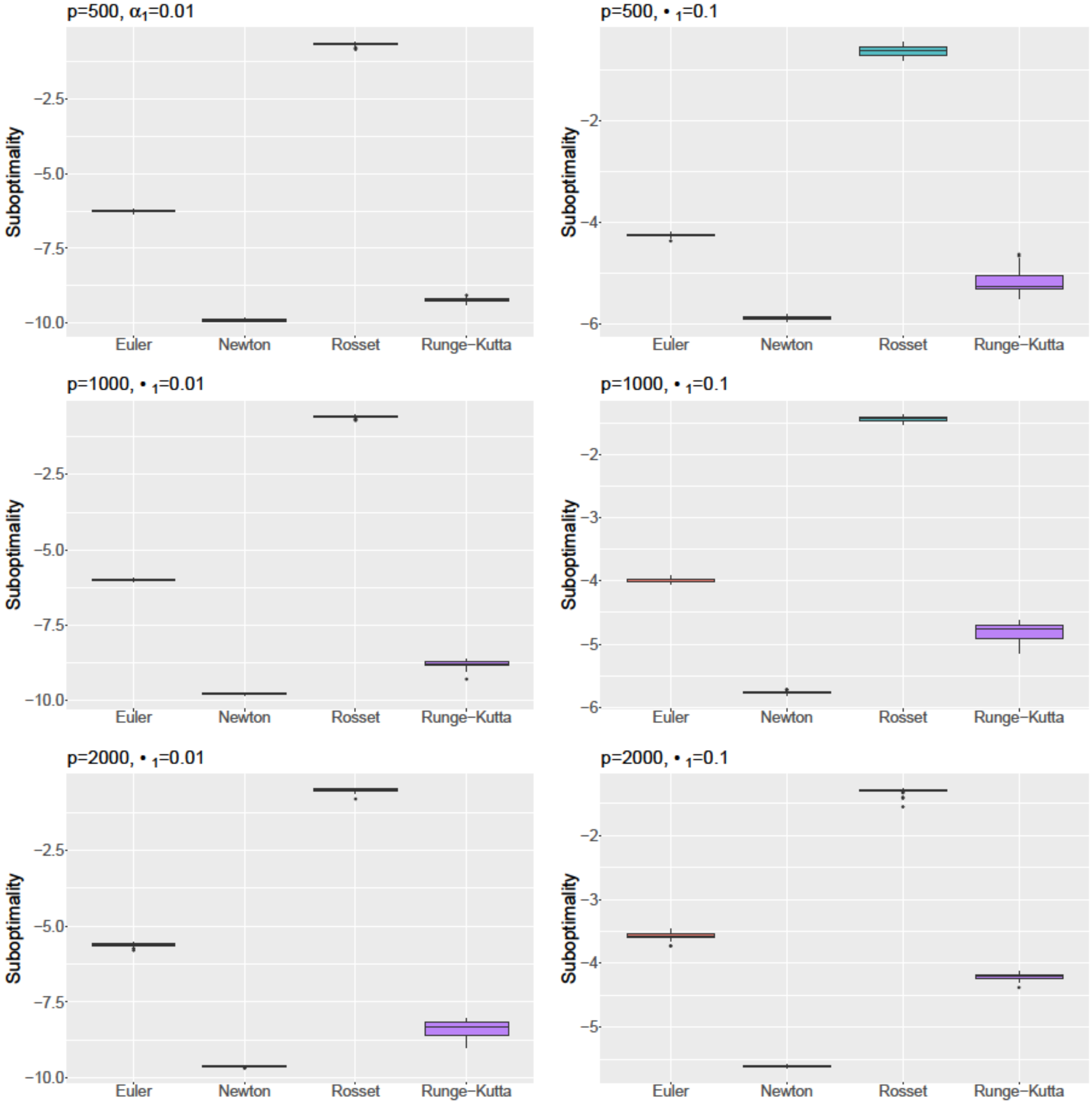


Figure 1: Suboptimalities $\sup_{0\leq t\leq 10}\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\}$ (in log scale) of the approximate solution paths generated by the proposed Newton method (Newton), the second-order Runge-Kutta method (Runge-Kutta), the Euler method (Euler), and the method of Rosset (2004) (Rosset) for ℓ_2-regularized logistic regression when the data is nonseparable.

suboptimality is small. This could be partially explained by the fact that the coordinate descent algorithms can usually be viewed as a type of methods that is between "first-order" and "second-order" method.

In summary, in terms of approximation error and computational efficiency, the Newton method and the second-order Runge-Kutta method both work quite well when the problem dimension is not too large or the desired suboptimality is small. For large-scale problems,

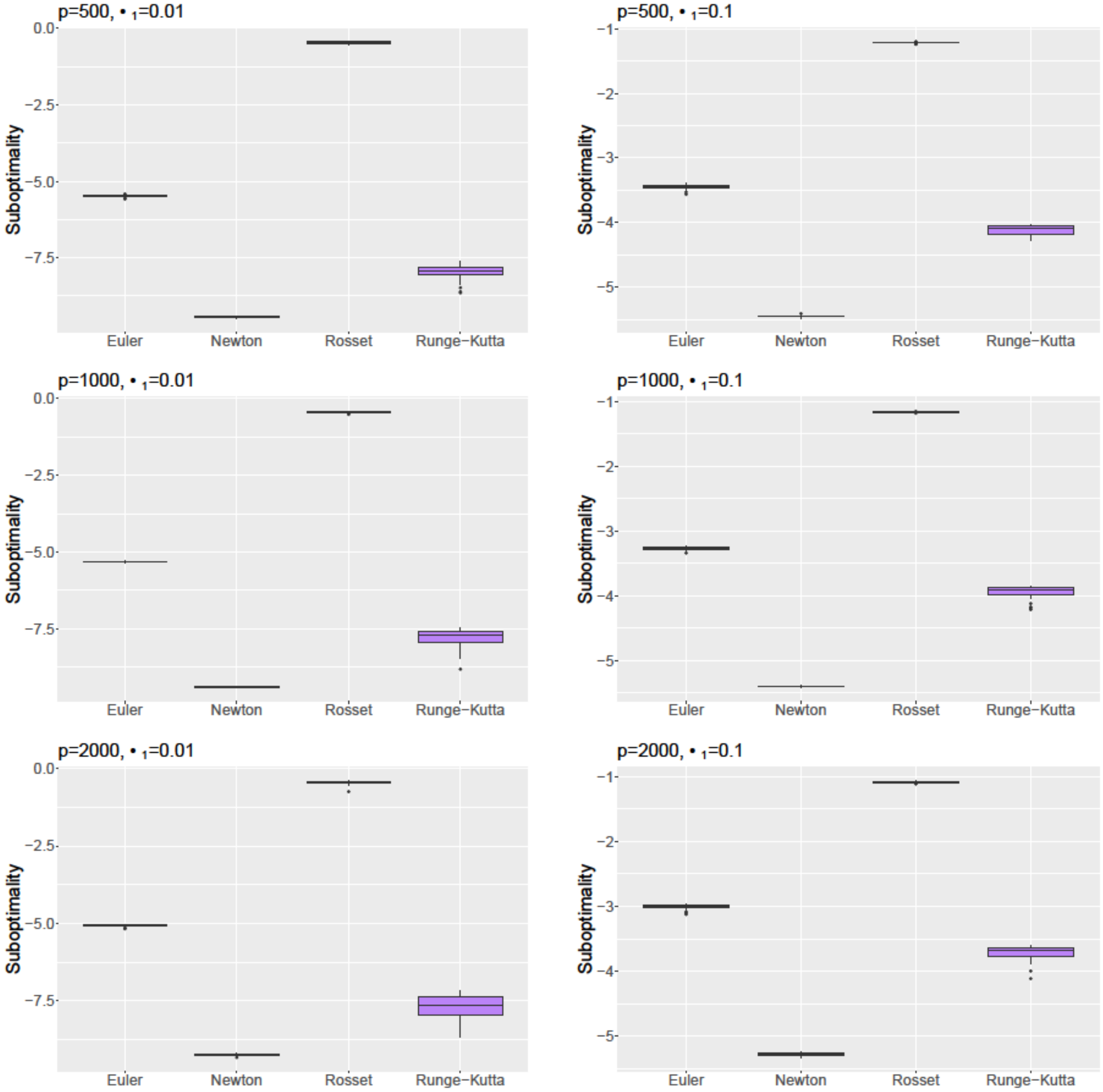
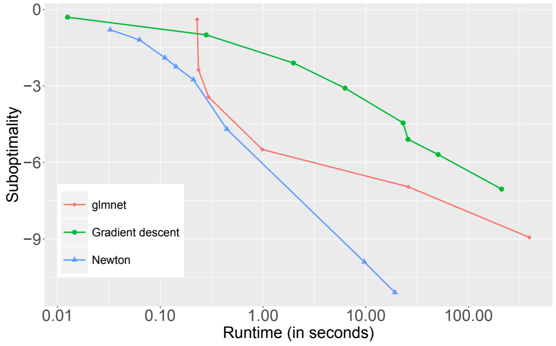


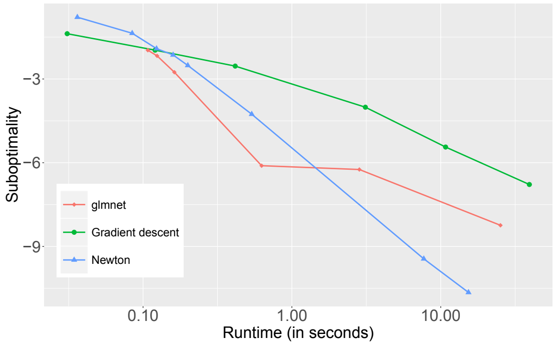
Figure 2: Suboptimalities $\sup_{0\le t\le 10}\{f_t(\tilde{\theta}(t)) - f_t(\theta(t))\}$ (in log scale) of the approximate solution paths generated by the proposed Newton method (Newton), the second-order Runge-Kutta method (Runge-Kutta), the Euler method (Euler), and the method of Rosset (2004) (Rosset) for ℓ_2-regularized logistic regression when the data is separable.

however, gradient descent method and glmnet seem to be more scalable, although glmnet produces solution paths with better suboptimality.

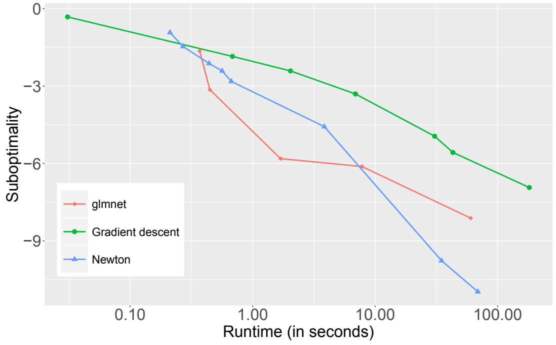
Lastly, we investigate how the initial step size of various solution path algorithms would affect their statistical performances. As we have argued before, the initial step size determines the approximation error. To assess the accuracy of the approximation to the true statistical risk, we consider a generative model for logistic regression. Specifically, we first generate the predictors $X_1, \ldots, X_n \in \mathbb{R}^p$ from normal distribution $N_p(0, I_{p\times p})$. Given prep=500, nonseparable



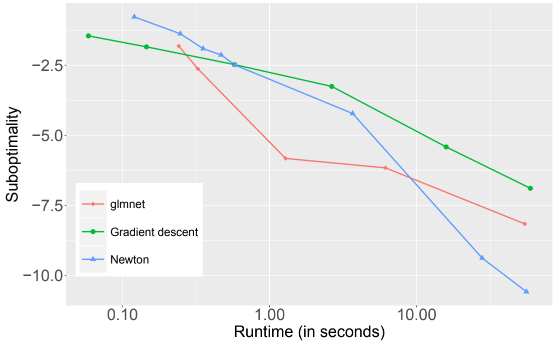
p=500, separable



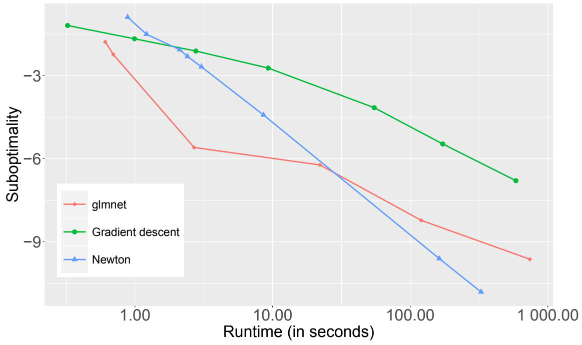
p=1000, nonseparable



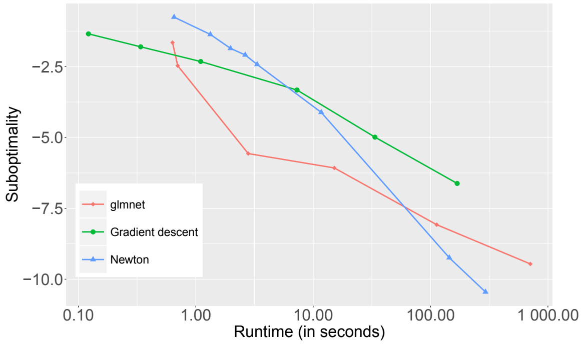
p=1000, separable



p=2000, nonseparable



p=2000, separable



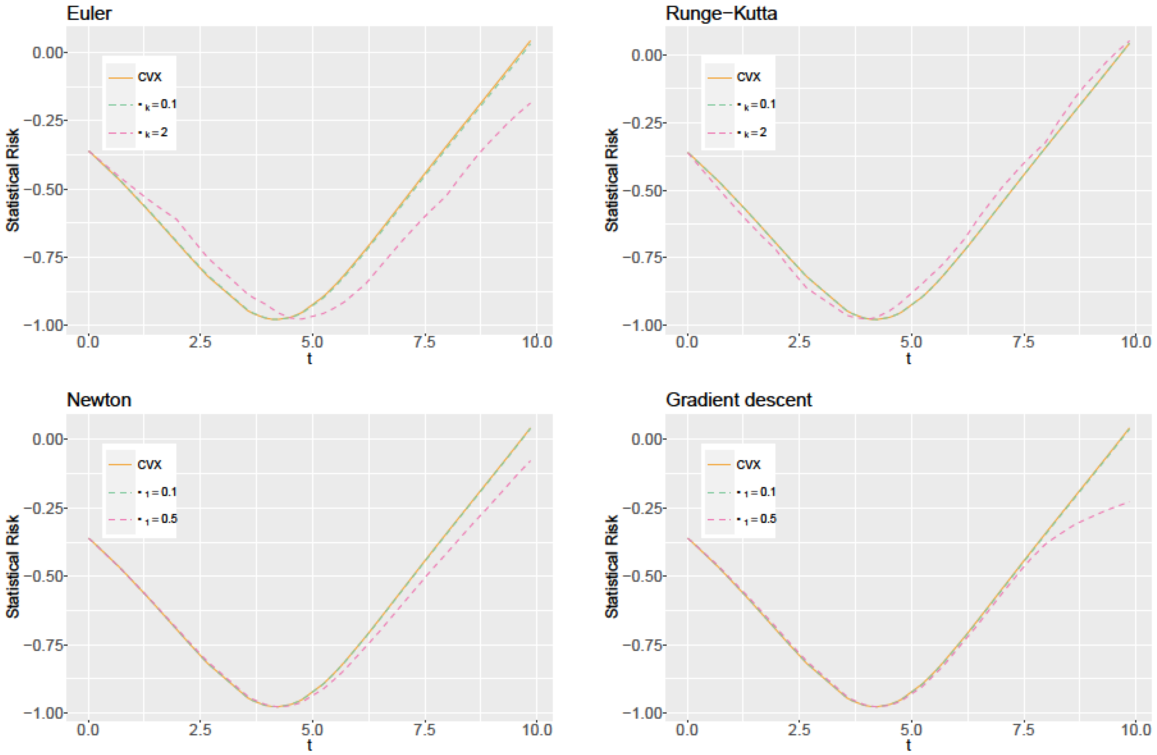


Figure 4: Approximate risk curve $\log_{10}(R(\tilde{\theta}(t), \theta))$ of the proposed algorithms applied to ℓ_2-regularized logistic regression when problem dimension is $(n, p) = (500, 100)$. The CVX (orange) curve denotes the true risk curve $\log_{10}(R(\theta(t), \theta))$ with $\theta(t)$ computed using the CVX solver. For algorithms with constant step size (Euler and Runge-Kutta), α_k denotes the step size; while α_1 denotes the initial step size for Newton and gradient descent method.

Note that the statistical risk for the exact solution path $\theta(t)$ is $R(\theta(t), \theta)$, which we refer to as the true risk curve (as a function of t). Here, we calculate the exact solution path $\theta(t)$ using CVX (Grant and Boyd, 2014, 2008). Again, the goal is to see the impact of the initial step size on how close the approximate risk curve $R(\tilde{\theta}(t), \theta)$ is to the true risk curve $R(\theta(t), \theta)$.

Figures 4–6 plot the approximate risk curve $R(\tilde{\theta}(t), \theta)$ against the true risk curve (on a log scale) by varying the initial step sizes for the proposed methods. Note that under all scenarios, when the initial step size is 0.1 (i.e., $\alpha_1 = 0.1$), the approximate risk curves approximate the true risk curve quite well for all four methods. This seems to suggest that good approximation error leads to good approximation of the risk curve. As the initial step size increases, interestingly, we observe that Runge-Kutta continues to provide reasonable good results, suggesting that they are more tolerant of a large initial step size (see the results when $\alpha_k = 2$ for Runge-Kutta methods on Figures 4–6). On the other hand, the Newton method and the gradient descent method requires the initial step sizes to be much smaller to obtain reasonable risk curve approximation. That says, this does not necessarily imply that the Newton method is less efficient than the ODE-based methods, because the

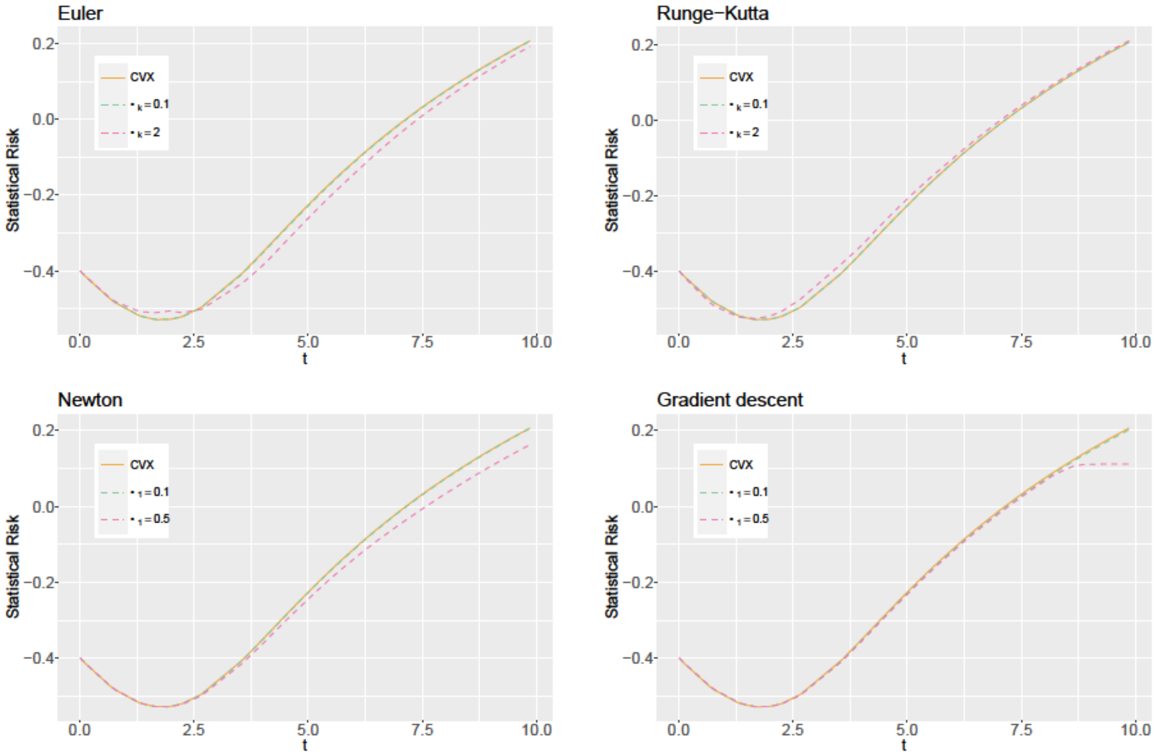


Figure 5: Approximate risk curve $\log_{10}(R(\tilde{\theta}(t), \theta))$ of the proposed algorithms applied to ℓ_2-regularized logistic regression when problem dimension is $(n, p) = (500, 500)$. The CVX (orange) curve denotes the true risk curve $\log_{10}(R(\theta(t), \theta))$ with $\theta(t)$ computed using the CVX solver. For algorithms with constant step size (Euler and Runge-Kutta), α_k denotes the step size; while α_1 denotes the initial step size for Newton and gradient descent method.

Newton method will adaptively increase step sizes while the ODE-based methods always fix their step sizes.

6. Discussion

In this article, we established a formal connection between ℓ_2-regularized solution path and the solution of an ODE. This connection provides an interesting algorithmic view of ℓ_2 regularization. In particular, the solution path turns out to be similar to the iterates of a hybrid algorithm that combines the gradient descent update and the Newton update. Moreover, we proposed various new path-following algorithms to approximate the ℓ_2-regularized solution path. Global approximation-error bounds for these methods are also derived, which in turn suggest some interesting schemes for choosing the grid points. Computational complexities are also derived using the proposed grid point schemes.

One important aspect we did not touch on is the statistical properties of ℓ_2-regularized solution path, which has been studied extensively in the literature (see, e.g., Dobriban and Wager, 2018, and references therein). Interestingly, Ali et al. (2019), in the context of least squares regression, connects the statistical properties of gradient descent iterates to that

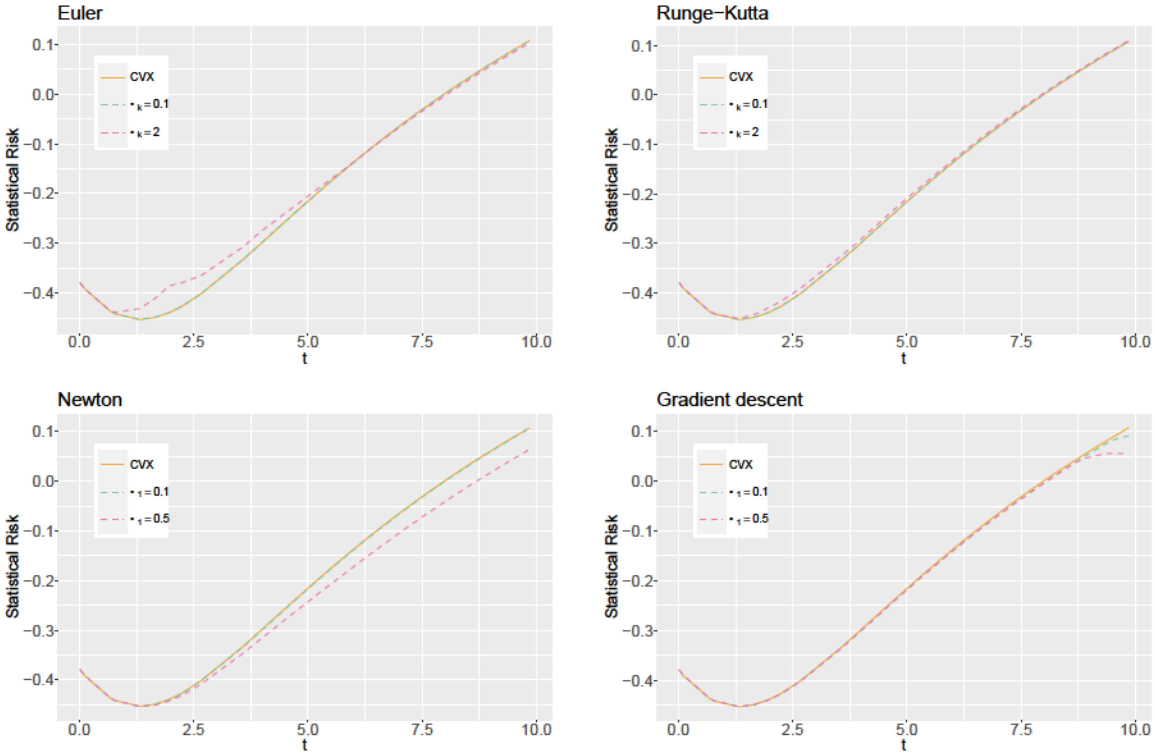


Figure 6: Approximate risk curve $\log_{10}(R(\tilde{\theta}(t), \theta))$ of the proposed algorithms applied to ℓ_2-regularized logistic regression when problem dimension is $(n, p) = (500, 1000)$. The CVX (orange) curve denotes the true risk curve $\log_{10}(R(\theta(t), \theta))$ with $\theta(t)$ computed using the CVX solver. For algorithms with constant step size (Euler and Runge-Kutta), α_k denotes the step size; while α_1 denotes the initial step size for Newton and gradient descent method.

of ridge regression solution path. In particular, they show that the statistical risk of the gradient descent path is no more than 1.69 times that of ridge regression, along the entire path. Motivated by our proposed homotopy method based on damped gradient descent updates (9), it would be interesting to investigate whether a damped version of gradient descent algorithm would enjoy a more favorable statistical risk compared to regular gradient descent. Further investigation is necessary.

Acknowledgments

We would like to thank the Associate Editor and reviewers for their insightful comments and encouragement to revise our paper. The feedback substantially improved the paper. We would also like to acknowledge support for this project from the National Science Foundation (DMS-17-12580, DMS-17-21445 and DMS-20-15490).

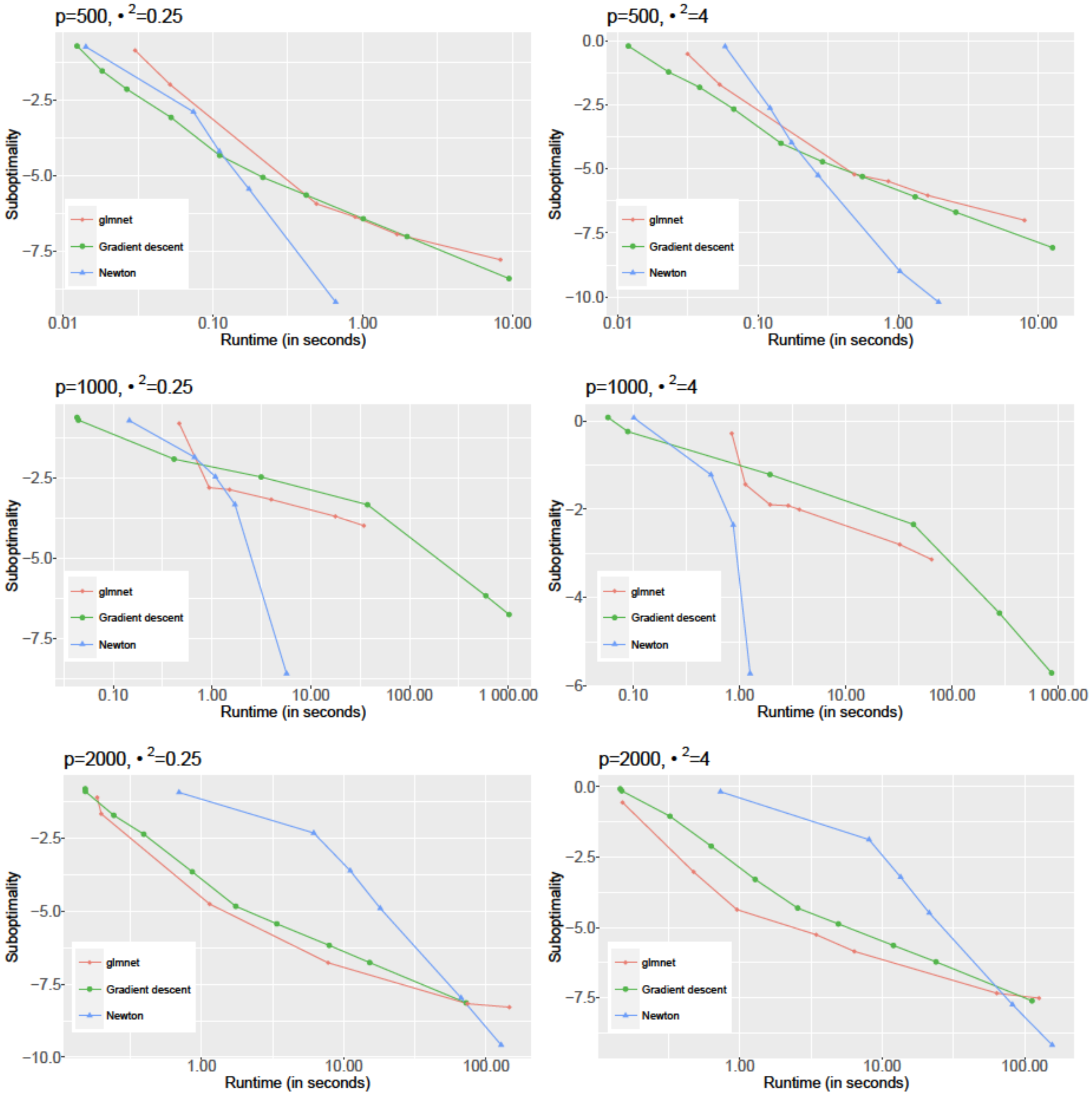


Figure 7: Runtime v.s. suboptimality for the proposed Newton method, gradient descent method, and glmnet under six different scenarios, when applied to ridge regression.